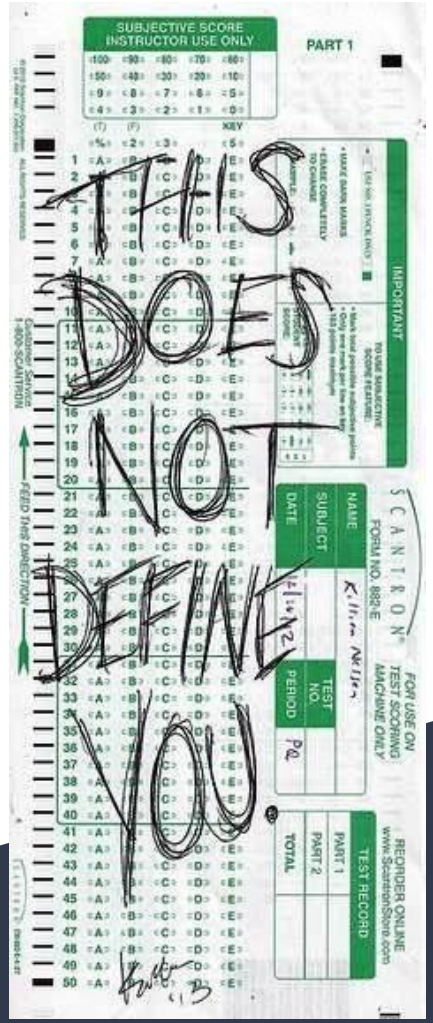
Project 1 - Standardized Test Analysis

Janet, Matthew, Clement



Standardized tests raise tons of questions

Standardized tests are controversial:

Are they the right selection tool to get into college?

If yes, which one should be taken? SAT or ACT?

Based on articles and datasets, we have chosen to focus on SAT 2017-2018-2019 data per state

Problem statement:

"Does looking at SAT scores averages per state paint a correct picture of academic performance for a given state?"

Part 2 - Data import and Cleaning

First part of an EDA is to do some data cleaning

Ensuring df integrity:

- existing dimensions,
- missing or wrong values
- replacing, deleting (cleaning) what needs to be
- renaming columns

Useful methods used :

pd.read_csv (to create dataframes)

.loc (to filter df)

.replace (to replace a certain character in a given field)

.apply(to apply a function to a dimension)

.rename(columns ={x,y})

.concat (to “stack” homogenous df together)

.drop(columns=[‘xyz]) to remove columns from df

Import and cleaning output: 1 containing 3 years of data

Feature	Type	Dataset	Description
state	object	sat_three_years	states where SAT was taken
score_read_write	int	sat_three_years	score obtained in reading and writing
score_math	int	sat_three_years	score obtained in math
score_total	int	sat_three_years	sum of reading_writing and math
participation_percent	float	sat_three_years	% of the student population who have taken the test
year	int	sat_three_years	year of the test

Exploratory Data Analysis

Large part of addressing a problem statement is the Exploratory Data Analysis (EDA)

In this phase, we discover the major characteristics contents in our data with help of descriptive summary statistics and graphical representation.

Useful methods used:

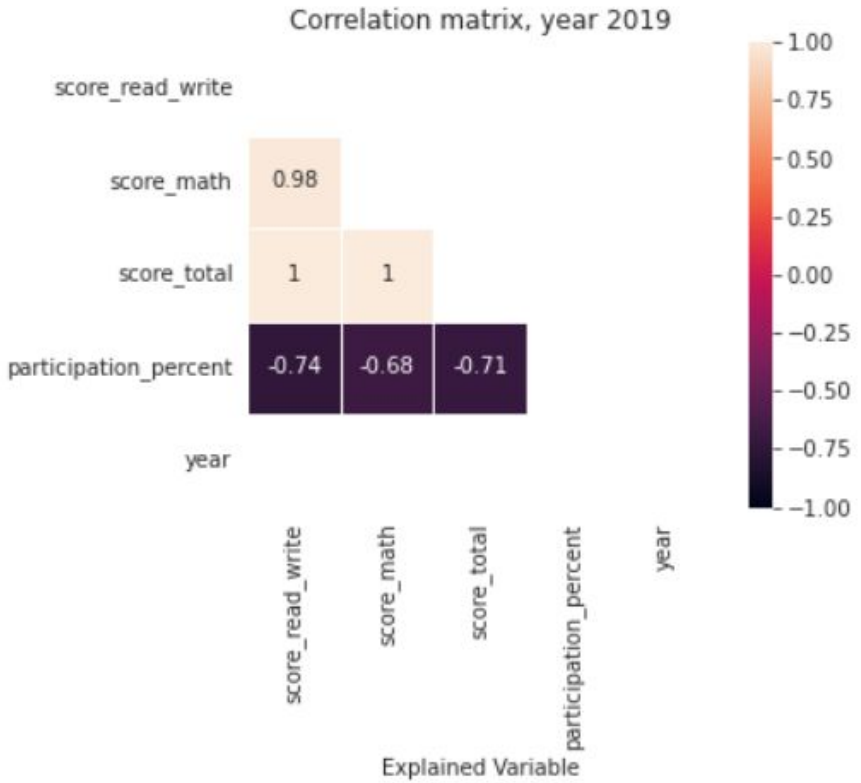
.groupby (to summarize mean, sum and variance in a pivot-table fashion)

.corr (to see correlation matrix in a table form)

.sort_values

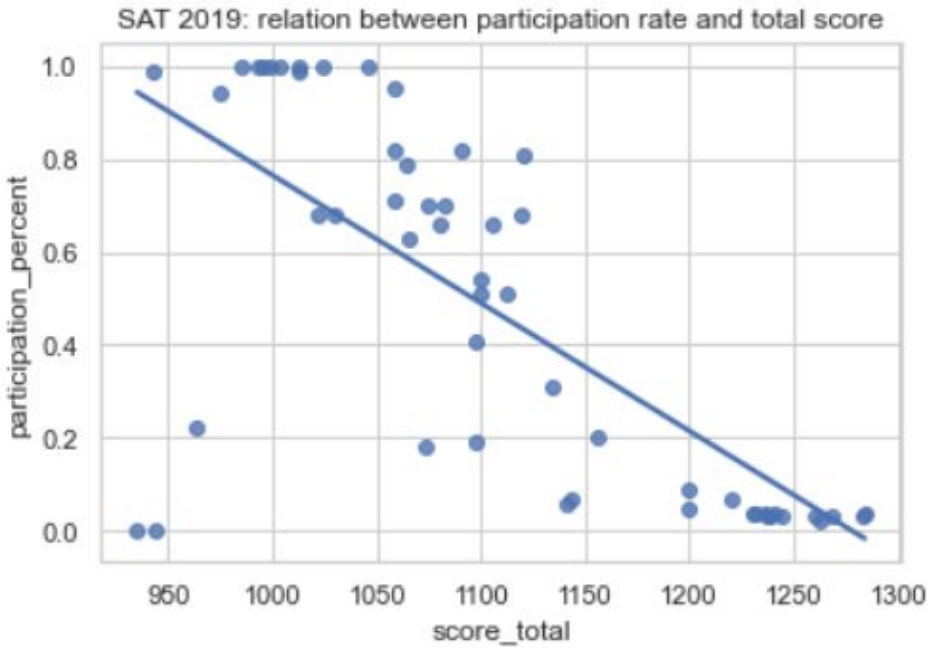
Visualization

Correlation matrix showing a strong negative correlation (>|0.5|) between participation percentage and score_total



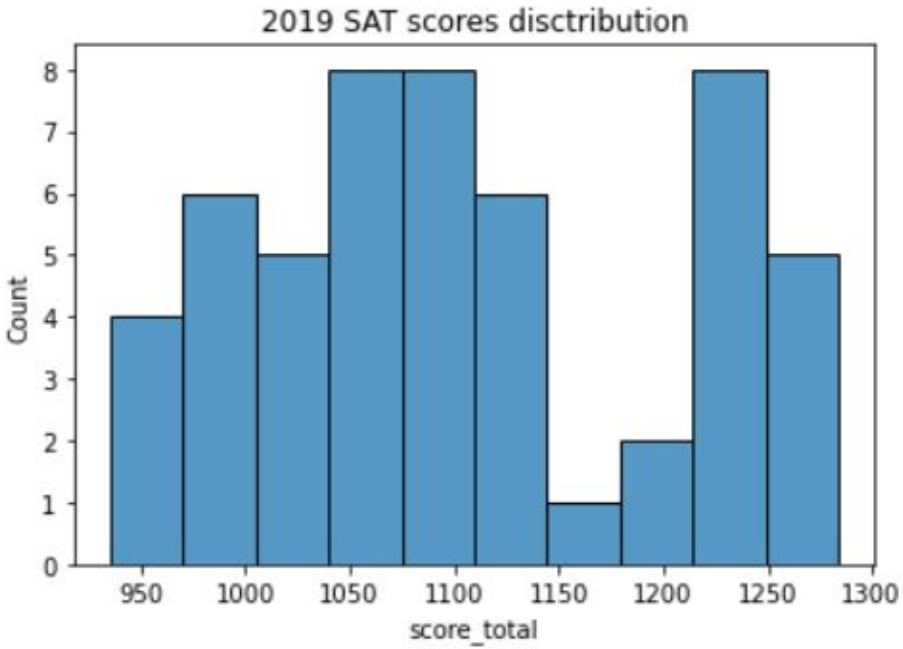
Visualization

Year 2019 scatterplot showing SAT total score vs participation rate



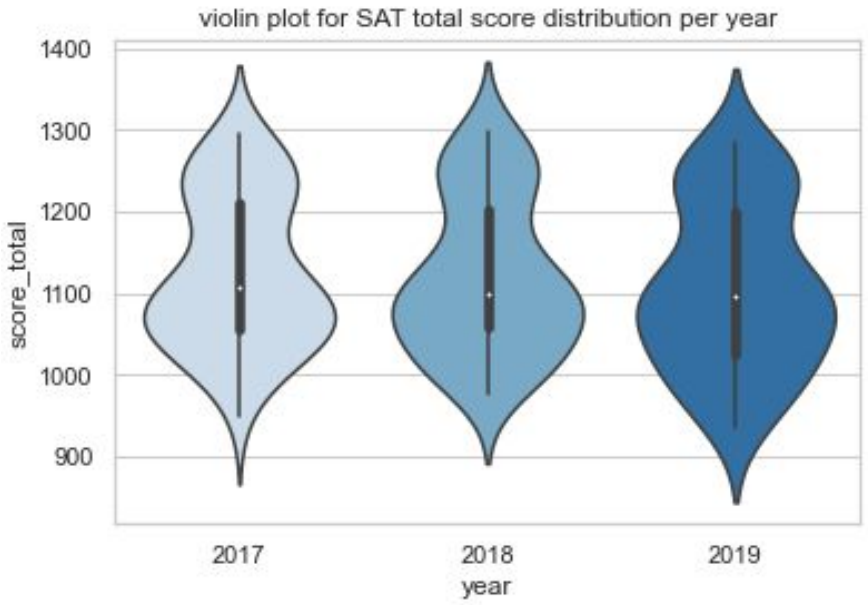
Visualization

SAT 2019 scores distribution shows 2 groups of states (not “normally distributed”)



Visualization

Year 2019 has a larger spectrum of scores due to the addition of 2 states to the dataset



Conclusion

Problem statement:

"Does looking at SAT scores averages per state paint a correct picture of academic performance for a given state?"

SAT Score results seem heavily correlated to the % of participation in each state. this is verified for the last 3 years.

From several articles, this is explained by some states enforcing SAT before graduating from high school, and some don't.

This warrant a hypothesis testing to validate the significance of the relationship we observe.

We cannot answer the problem statement without further hypothesis testing.